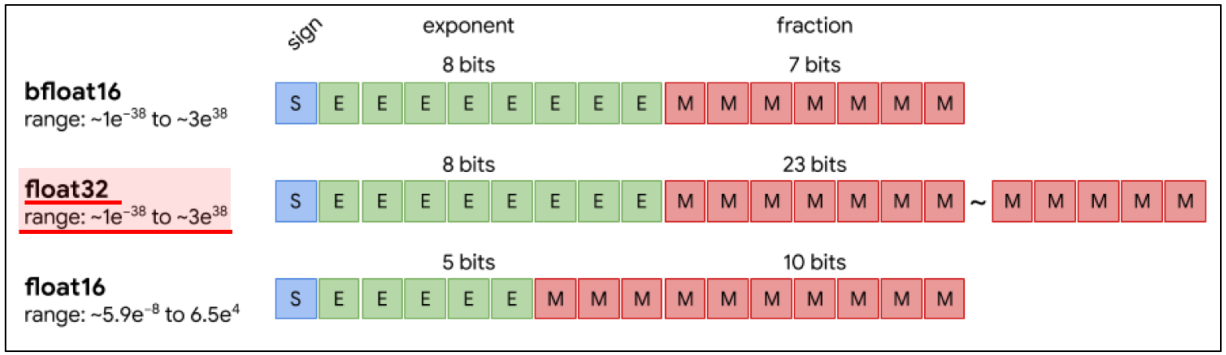
Exhibit 8

Exhibit A

U.S. Pat. No. 9,218,156

Claim 7

'156 PATENT	INFRINGEMENT EVIDENCE
7. A device comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and **for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input;** at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	• "Each TPU core has scalar, vector, and matrix units (MXU). The MXU provides the bulk of the compute power in a TPU chip. Each MXU is capable of performing 16K multiply-accumulate operations in each cycle. While **the MXU inputs and outputs are 32-bit floating point values**, the MXU performs multiplies at reduced bfloat16 precision. Bfloat16 is a 16-bit floating point representation that provides better training and model accuracy than the IEEE half-precision representation." *https://cloud.google.com/tpu/docs/system-architecture* • "The following figure shows three floating-point[] formats • **fp32 - IEEE single-precision floating-point** • fp16 - IEEE half-precision floating point • bfloat16 - 16-bit *brain floating point*" *https://cloud.google.com/tpu/docs/bfloat16*  *Id.* • "Because general-purpose processors such as CPUs and GPUs must provide good performance across a wide range of applications, they have evolved myriad sophisticated, performance-oriented mechanisms. As a side effect, the behavior of those processors can be difficult to predict, which makes it hard to guarantee a certain latency limit on neural network inference. In contrast, TPU design is strictly **minimal and deterministic** as it has to run **only one task at a time: neural network prediction**. You can see its simplicity in the floor plan of the TPU die." *https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu (emphasis in orig.)* • "In mathematics, computer science and physics, a **deterministic** system is a system in which **no randomness** is involved in the development of future states of the system. A **deterministic model will thus always produce the same output from a given starting condition or initial state**." *https://en.wikipedia.org/wiki/Deterministic_system* • For each of the possible valid inputs to the multiplication operation performed by the multipliers within the MXU, Singular has computed the result and compared it to the result of an exact mathematical calculation performed on the same inputs. The results of this test showed that for **more than 10%** of the possible valid inputs, the numerical value represented by the output signal of each MXU multiplier differs by **more than 0.2%** from the result of an exact mathematical calculation performed on the same inputs.